一种面向表情识别的 ROI 区域二级投票机制 *

文元美，欧阳文，凌永权

(广东工业大学 信息工程学院，广州 510006)

摘 要：针对如何更有效地使用卷积神经网络从训练图像中学习到的分布式特征进行研究，提出了一种面向人脸表情识别的 ROI 区域二级投票机制。首先将图像划分成一系列感兴趣区域（ROI）图像输入到卷积神经网络中进行训练；然后将测试图像的 ROI 图像输入到卷积神经网络中，统计所有 ROI 图像的判别结果；最后采用二级投票机制确定测试图像的最终类别，得到最终判别结果。此外，针对卷积神经网络不能从人脸图像中学习到旋转等空间位置信息，引入了 STN(spatial transformer network)网络，提高算法在解决复杂情况下的表情识别问题的能力。实验表明，ROI 区域二级投票机制能够更有效地使用卷积神经网络从训练图像中学习到的分布式特征，比直接使用 ROI 图像进行投票的方法准确率提升了 1.1%，引入 STN 网络能够有效提升卷积神经网络的鲁棒性，比未引入 STN 网络的方法准确率提升了 1.5%。

关键词：卷积神经网络；表情识别；STN 网络；二级投票机制

中图分类号：TP391.4　　doi: 10.3969/j.issn.1001-3695.2018.03.0189

Expression-oriented ROI region secondary voting mechanism

Wen Yuanmei, Ouyang Wen, Ling Yongquan

(*School of Information Engineering Guangdong University of Technology*, *Guangzhou* 510006, *China*)

Abstract: Aiming at the problem of how to more efficiently use the distributed features that convolutional neural network have learned from training images, this paper proposed a ROI(regions of interest) region secondary voting mechanism for facial expression recognition. Firstly, it divided into the image a series of ROI images, and input it into the convolutional neural network for training. Then, it input into the ROI images of the test image the convolutional neural network, getting all ROI images' results. Lastly, it used the secondary voting mechanism to determine the final category of test image. In addition, aiming at the problem of convolutional neural network cannot learn spatial position information such as rotation, this paper introduced the STN (spatial transformer network) to make convolutional neural network useful in complex condition. Experiments show the ROI region secondary voting mechanism can more effectively use the distributed features which learned by convolutional neural network, compared with the method of voting directly using ROI images, the accuracy is increased by 1.1%. The introduction of STN can effectively improve the robustness of convolutional neural network, compared with non-introduced STN networks, the accuracy is increased by 1.5%.

Key words: convolutional neural network; expression recognition; STN network; secondary voting mechanism

0　引言

人脸表情是由人眼睛、鼻子、嘴巴、眉毛等处的肌肉形变产生的，是人类情感交流最有力、最自然、最直接的手段之一，能够正确反映人当前所处的状态。人脸表情识别能让计算机识别人的表情，并根据表情所反映的信息为人类提供更人性化的服务，在安全驾驶、人机交互、测谎等诸多方面有广泛的应用，因此成为计算机视觉的研究热点之一[1~3]。

自从 Krizhevsky 等人[4]利用卷积神经网络在 ILSVRC-2012 图像识别竞赛取得比手工特征更好的效果，卷积神经网络引起了广泛的关注，学者们对基于卷积神经网络的表情识别进行了一系列的探索与分析[5,6],比如 Hamester 等人[7]提出了一种由标准 CNN 通道与 CAE 通道构成的双通道卷积神经网络用于表情识别；Liu 等人[8]将卷积神经网络提取的特征与手工提取的 CBP 特征(centralized binary patterns) 相结合，使用 SVM 分类器进行分类；Meng 等人[9]在卷积神经网络中提取的表情特征中融合

收稿日期：2018-03-12；**修回日期：**2018-05-02　**基金项目：**国家自然科学基金资助项目（61372173，61671163）；2017 年中央财政支持地方高校建设项目

作者简介：文元美（1968-），女，湖北荆州人，副教授，博士，主要研究方向为智能信息处理（ym0218@gdut.edu.cn）；欧阳文（1994-），男，湖南郴州人，硕士研究生，主要研究方向为模式识别、深度学习；凌永权（1973-），男，中国香港人，教授，博士，主要研究方向为最优化信号处理与时频分析.

了人身份特征，提升了表情识别地准确率。但在这些方法中，卷积神经网络都是直接学习表情图像的全局特征，没有充分挖掘局部表情的分布式特征，引导卷积神经网络关注表情变化的重点区域。

ROI 区域即感兴趣区域。对于不同的计算机视觉任务，重点关注的区域也会不同。设定 ROI 区域，可以主动引导算法关注重点区域，从而达到提升识别精度、加快处理速度的功能。表情识别中，引入 ROI 区域，可以引导卷积神经网络关注表情相关的重点区域，从而提高表情识别的精度。已有研究者在基于卷积神经网络的表情识别中引入了 ROI 区域，比如 Vo 等人[10]分别使用全局图像与局部图像训练卷积神经网络，分别学习图像的全局信息和局部信息，在测试时利用局部图像测试的结果对全局图像的概率分布进行调整；孙晓等人[11]提出将人脸划分为一系列 ROI(regions of interest)区域输入到卷积神经网络中进行训练，在测试时利用 ROI 图像进行投票，选取票数最多的类别作为结果。上述方法在训练与测试中使用了 ROI 图像，同时也证明了在测试阶段使用 ROI 图像进行辅助判别有利于提升卷积神经网络的识别精度。

文献[11]中采用 ROI 投票进行辅助判别，较传统的基于卷积神经网络的表情识别方法取得了更好的效果。但是由于局部 ROI 图像包含的信息较少，导致 ROI 区域容易产生误判。为了充分使用卷积神经网络在训练阶段学习到分布式表达特征，同时降低由于 ROI 区域图像包含信息量太少对判别结果的影响，本文提出了面向表情识别的 ROI 区域二级投票机制，通过对全局图像赋予更大的影响因子，降低局部 ROI 图像对判别结果的影响。另外，为了解决文献[11]中提出的表情识别中卷积神经网络不具有旋转不变性问题，本文将 STN 网络(spatial transformer network)[12]引入了表情识别中，使卷积神经网络能够学习到表情图像的空间位置信息，提升系统的鲁棒性。

1　模型改进方法

本文在文献[11]模型的基础上对 ROI 图像的辅助判别方法以及模型的旋转不变性进行了研究。接下来，对上述两个方面的研究分别进行介绍。

1.1　ROI 辅助判别方法研究

在卷积神经网络的训练阶段引入 ROI 图像，不仅可以扩充数据集，防止网络过拟合，又可以使卷积神经网络学习到表情图像的分布式表达特征。为了充分使用卷积神经网络训练阶段学习到的分布式表达特征，同时减少由于局部 ROI 图像包含信息过少而导致的判别时容易产生误判的问题，本文在文献[11]的基础上提出了 ROI 区域二级投票机制，提升表情识别的精度。

本文参考文献[11]，根据人脸结构，通过分割、遮挡、翻转、中心聚焦处理，设置 9 个不同的 ROI 区域。分割处理时重点关注眼睛、鼻子、嘴巴区域的变化，提取眼睛、鼻子、嘴巴区域图像，得到四幅 ROI 图像，分别记为 ROI0、ROI1、ROI2、ROI3；遮挡处理时分别遮挡脸的上半部分与下半部分，得到两幅 ROI 图像，分别记为 ROI4、ROI5；翻转处理考虑了拍摄的角度不同，将图像进行水平翻转，得到一幅 ROI 图像，记为 ROI6；中心聚焦去除头发等噪声对表情的影响，聚焦人脸表情的重点区域，得到一幅 ROI 图像，记为 ROI7。处理得到的 8 幅 ROI 图像加上初始图像(记为 ROI8)一共得到 9 幅 ROI 图像，9 幅 ROI 图像如图 1 所示。



图 1　ROI 区域图像

通常，中性表情的眼睛、嘴巴、鼻子等部位没有特别变化；高兴表情具有的特征为嘴角张大或者上扬、眼睛变细、鼻翼上翘；悲伤表情具有的特征为眉毛眼角向下倾、嘴巴张大或者嘴角向下；愤怒表情具有的特征为眉毛上竖、嘴角下扣、眉头紧锁、鼻孔上翘，有时伴随着嘴巴张开；惊讶表情具有的特征为张大嘴、瞪大眼，同时眉毛上扬。但是每种表情都具有一定幅值，不同幅值表情的表达形式也会不一样。例如有时表达高兴情感时嘴巴张开幅度比较大，若将嘴巴部分的 ROI 图像提取出来，此部分 ROI 图像与惊讶表情的在嘴巴处的 ROI 图像相似度比较高，因此若直接对此 ROI 图像进行判别，容易将此处的 ROI 图像误判为惊讶。因此，在投票判决时，应适当减少包含局部信息的 ROI 图像对最终结果的影响，提升具有全局信息的 ROI 图像对最终判别结果的影响。

本文受决策树多级决策思想的启发，提出了二级投票机制。对于决策树来说，通常采用信息增益来进行划分属性的选择，信息增益计算方法为

$$Ent(D) = -\sum_{k=1}^{|y|} p_k \log_2 p_k$$

$$Gain(D,a) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v)$$

其中：D 为样本集；V 表示样本具有的某一属性。信息增益越大，该属性的影响越大，则优先采用该属性设置节点。因此，本文为了提升具有全局信息的 ROI 图像在判别时的影响力，优先采用具有全局信息的 ROI 图像设置了判别节点。假设 V_i 为表情类别，其中 $i \in [0,4]$，W_j 为 W 中第 j 幅 ROI 图像，$j \in [0,8]$，W_{j1}、W_{j2} 为 W 中具有全局信息的 ROI 图像，卷积神经网络对 W_{j1}、W_{j2} 的判别结果为 V_{i1}、V_{i2}，卷积神经网络对 W 中所有图像的判别结果为 I，经本文提出方法辅助判别后输出的结果为 O，因此，本文提出的方法为：

$$
\begin{aligned}
& Input : I \\
& Output : \mathrm{O} \\
& if\ \ V_{i1} = V_{i2} \\
& \quad O <- V_{i1} \\
& else \\
& \quad \mathrm{C} <- frequent\left(V_{i(0 \le i \le 4)}\ in\ range(I)\right) \\
& \quad t <- \max(C) \\
& \quad \mathrm{O} <- find\left(t \text{ corresponding to } V_i\right)
\end{aligned}
$$

在划分的 9 个 ROI 区域图像中，ROI6 与 ROI8 包含了完整的表情信息，因此，在利用 ROI 图像进行投票时，提高 ROI6 与 ROI8 对判别结果的影响力，同时降低其他 ROI 图像对判别结果的影响。因此，本文提出的二级辅助判别的具体步骤如下：

a)将测试图像划分为一系列 ROI 区域图像。

b)将划分得到的 ROI 图像输入到训练好的卷积神经网络中，统计每一个 ROI 图像的判别结果。

c)比较判断 ROI6 与 ROI8 的判别结果是否一致。若一致，则将 ROI6 与 ROI8 的判别结果归并为测试图像的判别结果；若不一致，则利用 ROI-KNN 方法进行投票，选取票数最多的结果在线归并为测试图像结果。

1.2　旋转不变性研究

在实际应用中，人脸图像通常会有一定旋转角度，且拍摄的人脸图像也会有拍摄角度的远近之分。因此，提高系统对旋转、缩放图像的处理能力，可以进一步提升系统的实用性。

针对卷积神经网络不能学习图像的空间信息如旋转、缩放等问题。本文在原卷积神经网络结构的基础上，引入了 STN 网络用来解决模型不具有旋转不变性的问题。

STN 网络由谷歌 Jaderberg 等人[12]于 2015 年提出。STN 网络由 Localisation Network、Grid generator、Sampler3 个模块组成，STN 网络结构如图 2 所示：

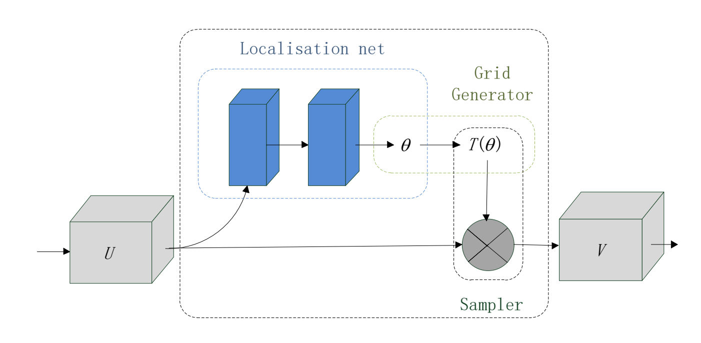


图 2　STN 网络结构

本文使用 STN 网络学习表情图像的位置信息的过程包含前向传播与反向调整两个阶段。

前向传播过程为：

a)将表情图像输入 Localisation Network 中，经过全连接层输出变换参数 θ，假设得到的 θ 为

$$\theta=\begin{bmatrix}\theta_{11} & \theta_{12} & \theta_{13}\\ \theta_{21} & \theta_{22} & \theta_{23}\end{bmatrix}$$

b)将得到的变换参数 θ 输入到 Grid generator 模块中，在此模块中得到生成图像与原图像坐标对应关系 T_θ。此过程如下所示：

$$\begin{bmatrix}x_i^s\\ y_i^s\end{bmatrix}=\begin{bmatrix}\theta_{11} & \theta_{12} & \theta_{13}\\ \theta_{21} & \theta_{22} & \theta_{23}\end{bmatrix}\begin{bmatrix}x_i^t\\ y_i^t\\ 1\end{bmatrix}=T_\theta(G_i)$$

其中：(x_i^t, y_i^t) 为生成网格的坐标；(x_i^s, y_i^s) 原图像的坐标。

c)利用得到的坐标对应关系进行插值，将原图像坐标中的像素值依照得到的坐标对应关系在生成网格中进行双线性插值，得到生成图像。公式表达如下：

$$V_i^C=\sum_n^H\sum_m^W U_{mn}^C\max\left(0,1-\left|x_i^s-m\right|\right)\max\left(0,1-\left|y_i^s-n\right|\right)$$

d)将插值得到的生成图像输入到卷积神经网络中进行特征学习。

反向传播过程为

e)计算 $\dfrac{\partial V_i^C}{\partial U_{nm}^C}$，使得误差能够经 STN 网络之后继续向前传播。计算公式如下：

$$\frac{\partial V_i^C}{\partial U_{nm}^C}=\sum_n^H\sum_m^W\max\left(0,1-\left|x_i^s-m\right|\right)\max\left(0,1-\left|y_i^s-n\right|\right)$$

假设卷积神经网络第一层输出的误差为 $loss$，经 STN 层输出误差为 $previous$，因此 STN 中误差反向传播的过程为

$$\frac{\partial loss}{\partial previous}=\frac{\partial loss}{\partial V_i^C}\bullet\frac{\partial V_i^C}{\partial U_{mn}^C}\bullet\frac{\partial U_{mn}^C}{\partial previous}$$

f)计算 $\dfrac{\partial V_i^C}{\partial x_i^s}$ 与 $\dfrac{\partial V_i^C}{\partial y_i^s}$，使得能够通过 V_i^C 反向调整变换变换参数 θ。$\dfrac{\partial V_i^C}{\partial x_i^s}$ 计算公式如下，$\dfrac{\partial V_i^C}{\partial y_i^s}$ 与 $\dfrac{\partial V_i^C}{\partial x_i^s}$ 计算过程相似。

$$\frac{\partial V_i^C}{\partial x_i^s}=\sum_n^H\sum_m^W U_{mn}^C\max\left(0,1-\left|x_i^s-m\right|\right)\max\left(0,1-\left|y_i^s-n\right|\right)\begin{cases}0, & \left|m-x_i^s\right|\ge 1\\ 1, & \mathrm{m}\ge x_i^s\\ -1, & m<x_i^s\end{cases}$$

因此，STN 中网络中参数 θ 反向调整过程为

$$\frac{\partial V_i^C}{\partial\theta}=\begin{pmatrix}\dfrac{\partial V_i^C}{\partial x_i^s}\bullet\dfrac{\partial x_i^s}{\partial\theta}\\[2ex] \dfrac{\partial V_i^C}{\partial y_i^s}\bullet\dfrac{\partial y_i^s}{\partial\theta}\end{pmatrix}$$

2　实验

为了验证本文提出的 ROI 区域二级投票机制的有效性以及引入 STN 是否能够使模型具有旋转不变性，本文就采用本文提出的方法(记为 Ours+CNN)与 ROI-KNN 方法(记为 ROI-KNN+CNN)以及直接对图像分类的方法(记为 ROI+CNN)进行了对比实验。同时在具有旋转的样本情况下，对在卷积神经网络中引入 STN 网络与未引入 STN 网络进行了对比实验。

2.1　实验样本选取及评价指标

本文使用的数据集是孙晓等人[11]采集的混合 CK 数据集与互联网采集的 Wild 面部表情数据形成的新的数据集，该数据集包含 CK+(Extended CohnKanade)[13]数据集中的高兴、悲伤、惊讶、愤怒各 700 张混合互联网下载的上述各种表情的 200 张 Wild 图像，以及 900 张实验室状态下的中性图像。共 5 类，每类 900 张，一共有 4 500 张图像。测试集由互联网采集除中性外其他类别的各 300 张混合 300 张实验室状态的中性图像，共 5 类 1 500 张图像，称为数据集 I。在数据集 I 的基础上，孙晓等人[11]对数据集 I 中的每张图像通过切割、翻转、遮盖、中心

聚焦处理后得到 9 张 ROI 图像，如图 1 所示，共 5 类，每类 4 500 张图像，测试图像不做变化，称为数据集 II。为了研究注入旋转样本能否使系统具有旋转不变性，孙晓等人[11]对数据集 I 包含的正规数据进行旋转生成采样，与数据集 II 相混合，得到 5 类共 83 500 张图片，称为数据集 III。本文采用了数据集 II 与数据集 III 进行实验。

本文实验采用准确率（accuracy）作为评价指标，准确率计算公式如下：

$$\text{Accuracy}=\left(1-\frac{\text{样本分类错误总数}}{\text{样本总数}}\right)\times 100\%$$

2.2　卷积神经网络结构及参数设置

本文采用的卷积神经网络结构参考文献[11]，采用 9 层卷积神经网络，包括了 3 个卷积层、3 个最大池化层、1 个全连接层、1 个 dropout[14]层、1 个 softmax 层。网络结构如表 1 所示。

表 1　本文采用的卷积神经网络结构

层数	类型	输出特征图	卷积核尺寸	池化核尺寸
0	Input	32*32*1		
1	Conv1	30*30*64	3*3	
2	Pool1	15*15*64		2*2
3	Conv2	12*12*64	4*4	
4	Pool2	6*6*64		2*2
5	Conv3	2*2*128	5*5	
6	Pool3	1*1*128		
7	Full	1*1*300		
8	Droput	1*1*300		
9	Softmax	1*1*5		

网络的第一层为卷积层，卷积核大小为 3*3，输出 64 个 30*30 的特征图；第二层为池化层，池化核大小为 2*2，输出 64 个 15*15 的特征图；第三层为卷积层，卷积核大小为 4*4，输出 64 个 12*12 的特征图；第四层为池化层，池化核大小为 2*2，输出 64 个 6*6 的特征图；第五层为卷积层，卷积核大小为 5*5，输出 128 个 2*2 的特征图；第六层为池化层，池化核大小为 2*2，输出 128 个 1*1 的特征图；第七层为全连接层，输出 300 个特征值；第八层为概率为 0.5 的 dropout 层；第九层为 softmax 层。卷积层的激活函数采用 ReLu[15]函数。

本文的权值与偏置的初始值服从均值为 0、标准差为 0.1 的标准正态分布。训练过程中，每次随机从样本中选取 100 个样本，共进行 50 000 次随机采样。初始学习率为 0.01，动量为 0.09，每采样 5 000 次验证一次，验证过程中发现准确率不变或下降时，学习率下降一个数量级继续训练，学习率下降到 0.000 1 时不再变化。

2.3　实验步骤

2.3.1 ROI 辅助判别实验步骤

本文采用数据集 II 验证本文提出的 ROI 辅助判别方法，实验主要由卷积神经网络训练、ROI 图像测试以及 ROI 区域辅助判别三个环节组成。

在卷积神经网络的训练阶段，首先将划分好的 ROI 图像按照下式进行归一化，将归一化之后的图像输入到卷积神经网络中进行训练，得到训练好的卷积神经网络模型。式中：$train_image$ 是归一化之后的图像；$image$ 是原始图像。

$$train_image=(image-(255/2.0))/255$$

在 ROI 图像测试阶段，首先测试图像的 ROI 图像进行归一化，然后将归一化之后的图像输入到训练好的卷积神经网络当中，统计测试图像的 ROI 图像的判别结果。

在最终判别阶段，利用本文提出的方法对 ROI 图像测试阶段统计得到的结果进行处理，得到最终的判别结果。

2.3.2 旋转不变性研究实验步骤

本文采用数据集 III 进行旋转不变性的研究。首先将数据集 III 输入到卷积神经网络当中进行训练，训练完成之后将测试数据集输入到训练好的网络当中，记录测试的准确率。

接下来在卷积神经网络的第一层中引入 STN 网络，同样将数据集 III 输入到引入了 STN 的卷积神经网络当中进行训练，训练完成之后将测试图像输入到训练好的网络当中，记录测试的准确率。

3　实验结果与分析

3.1　ROI 辅助判别实验结果与分析

为验证本文方法的有效性，将本文提出的方法与 ROI-KNN 方法以及利用 ROI 图像进行数据增强的方法进行对比实验，实验结果如表 2 所示。

表 2　不同辅助判别方法准确率对比

方法	准确率
Ours+CNN	78%
ROI-KNN+CNN	76.9%
ROI+CNN	73.2%

由表 2 可知，在整体准确率上，采用本文提出的方法进行人脸表情识别准确率达到 78%，比 ROI-KNN+CNN 方法准确率提高了 1.1%，比 ROI+CNN 方法识别准确率提高了 4.8%。由此可见，本文提出的方法取得了最好的效果。

为了分析本文方法取得最好效果的原因，本文在实验中引入了混淆矩阵观察每个类别的分类情况。Ours+CNN 方法实验结果的混淆矩阵如图 3 所示，ROI-KNN+CNN 方法实验结果的混淆矩阵如图 4 所示，ROI+CNN 方法实验结果的混淆矩阵如图 5 所示。

混淆矩阵横轴为预测的类别，纵轴为实际类别。矩阵中从左到右依次是中性、高兴、悲伤、惊讶、愤怒。对角线中的数值表示每个类别的分类的准确率。

由混淆矩阵可知，本文提出的方法在自然、高兴、悲伤、惊讶、愤怒表情中的识别准确率分别为 0.98、0.71、0.56、0.89、0.76；ROI-KNN+CNN 方法在上述表情的识别准确率分别为 0.98、0.69、0.54、0.88、0.75；ROI+CNN 法在上述表情的识别

准确率分别为 0.96、0.66、0.49、0.83、0.68。

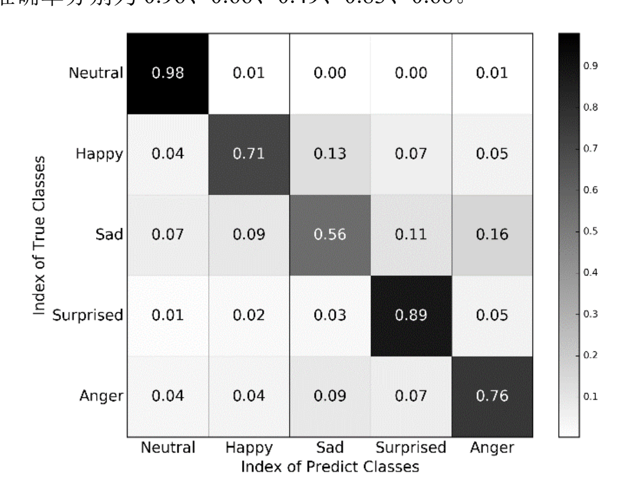


图 3　Ours+CNN

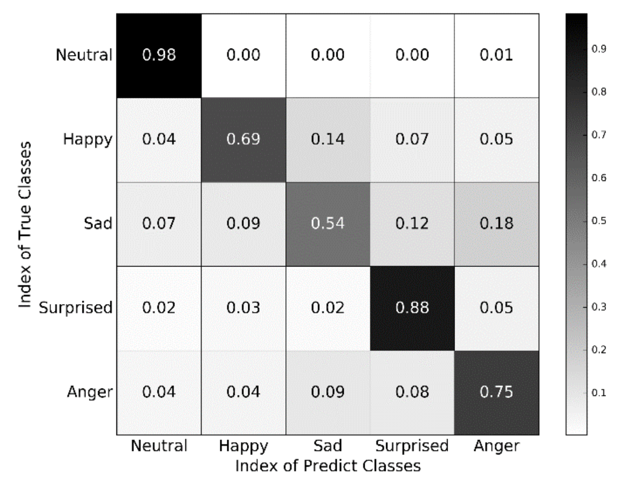


图 4　ROI-KNN+CNN

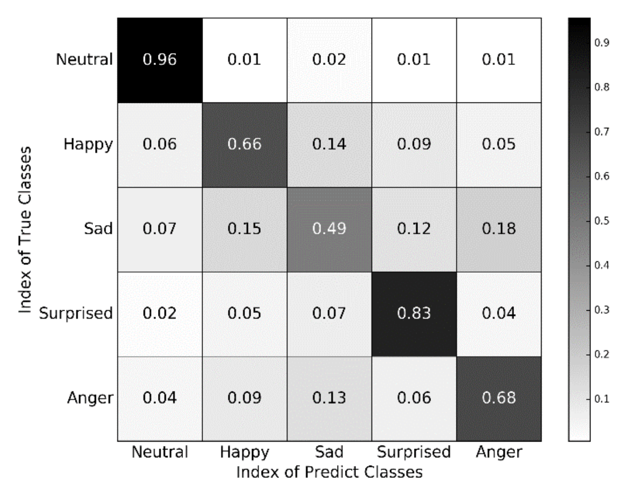


图 5　ROI+CNN

由此可见，本文提出的方法与 ROI-KNN 方法在各个表情都能取得比 ROI+CNN 更高的准确率，说明在测试阶段引入 ROI 图像进行辅助判别，充分使用网络从 ROI 图像中学习到的信息，能有效地提升系统对表情地判别能力。对于本文提出的方法与 ROI-KNN 方法，由混淆矩阵可知，本文提出的方法与 ROI-KNN 方法在各表情类别中准确率相接近甚至更高，在整体准确率上，本文提出的方法要比 ROI-KNN 方法准确率更高。

为了分析本文提出方法能够取得比 ROI-KNN 卷积神经网络取得更好效果的原因，本文挑选出本文分类正确、ROI-KNN 方法分类错误的样本，如图 6 所示，并对这些样本的 ROI 图像的类别概率分布进行可视化，如表 3 所示。由于情况相同，本文以图 6 以及图 6 对应的类别概率分布进行分析。

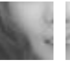
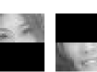
      

图 6　ROI-KNN 错分样本

表 3　ROI-KNN 错分样本类别概率分布

ROI 图像	类别概率分布				
	中性	高兴	悲伤	惊讶	愤怒
0	0.000	0.000	0.100	0.900	0.000
1	0.749	0.245	0.005	0.002	0.000
2	0.000	0.002	0.118	0.855	0.024
3	0.002	0.561	0.131	0.307	0.000
4	0.000	0.017	0.139	0.844	0.000
5	0.000	0.629	0.357	0.013	0.000
6	0.000	0.001	0.994	0.005	0.000
7	0.839	0.149	0.009	0.000	0.002
8	0.000	0.992	0.003	0.005	0.000

表 3 中的 ROI 图像分别对应于图 6 中从左到右的图像。对于每幅图像，通常都是选取类别概率分布中最大概率值对应的类别作为此 ROI 图像的类别。

表 3 中，ROI0、ROI2、ROI4 都在惊讶的表情中拥有最大概率值 0.900、0.855、0.844，因此有 3 幅 ROI 图像被判别为惊讶。同理，ROI3、ROI5、ROI8 在高兴表情的概率分别为 0561、0.629、0.992，在各类别概率分布中拥有最大值，因此这 3 幅 ROI 图像被判别为高兴，出现了多幅局部 ROI 图像产生误判的情况，此时导致 ROI-KNN 产生误判。将图像划分成 ROI 图像后，局部 ROI 图像包含的信息较少，这些包含局部信息的 ROI 图像与其他表情的 ROI 图像比较相似，容易将 ROI 图像判断为具有相似 ROI 图像的其他表情。当某测试图像中多个 ROI 图像被误判为具有相似 ROI 图像的其他表情时，容易出现误判的 ROI 图像与正确判断的 ROI 图像相等甚至超过正确判断的 ROI 图像的情况，使得 ROI-KNN 方法产生误判。而本文提出的 ROI 区域二级投票机制，首先将两幅具有完整信息的 ROI 图像的判断结果进行比较，当结果不一致时，再采用投票机制，确定最终的判决结果，此时可以在一定程度上降低包含局部信息的 ROI 图像对最终判决结果的影响，因此取得了更好的效果。

另外，由于 ROI+CNN 对图像进行判别时只需要对测试图像进行一次判别，而本文与 ROI-KNN 方法都需要对测试图像的所有 ROI 图像进行判别，所以需要进行 9 次判别。因此本文还针对引入辅助判别是否会增加判别时间进行实验。经实验得，判别 1 500 张图像，ROI+CNN 方法用了 0.125 s，引入辅助判别方法用了 0.680 s。可见，引入了辅助判别仅比未引入辅助判别方法多了 0.555 s，但是准确率确提升了 4.8%。因此，本文提出方法能够在略微增加判决时间的前提下，取得更好的判别结果。

3.2　旋转不变性研究实验结果与分析

本文利用数据集 III 设置了 6 组实验，其中 3 组实验引入了 STN，3 组未引入 STN。实验结果如表 4 所示。

表 4　引入 STN 实验结果对比

未引入 STN	准确率	引入 STN	准确率
ROI+CNN	73.5%	ROI+ CNN	75%
Ours+CNN	76.7%	Ours+ CNN	78.4%
ROI-KNN+CNN	74.8%	ROI-KNN+CNN	76.9%

由表 4 可知，在 ROI+CNN 实验中，引入 STN 网络进行表情识别准确率达 75%，比未引入 STN 网络的识别准确率提高了 1.5%；在 Ours+CNN 实验中，引入 STN 网络识别准确率达 78.4%，比未引入 STN 网络的识别准确率提高了 1.7%。在 ROI-KNN+CNN 实验中，引入 STN 网络识别准确率为 76.9%，比未引入 STN 网络准确率提升了 2.1%。因此，在训练样本中注入了旋转样本的情况下，引入 STN 网络能够提升表情识别的准确率。同时，由在未引入 STN 网络以及引入 STN 情况下，本文提出的二级投票机制的识别准确率都高于 ROI-KNN 方法，进一步证明了本文提出的二级投票机制的有效性。

为了进一步观察在表情识别任务中引入 STN 网络的效果，本文引入了混淆矩阵进行观察。由于 ROI+CNN 中未使用辅助判别方法，所以以 ROI+CNN 中引入 STN 网络与未引入 STN 网络生成的混淆矩阵为例进行分析。ROI+CNN 中未引入 STN 网络实验结果的混淆矩阵如图 7 所示，引入 STN 网络实验结果的混淆矩阵如图 8 所示。

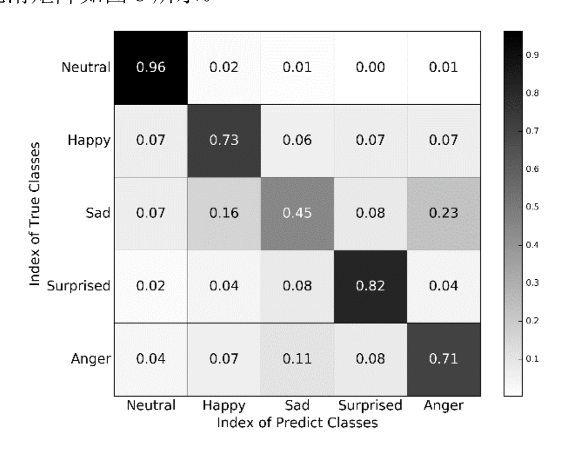


图 7　未引入 STN 网络

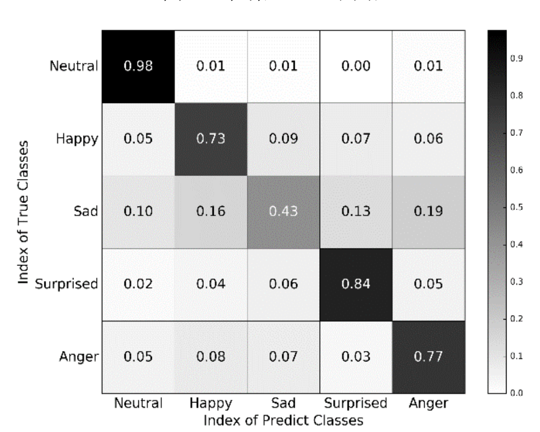


图 8　引入 STN 网络

由混淆矩阵可知，卷积神经网络中引入 STN 网络后，在自然、高兴、悲伤、惊喜、愤怒五种情感中的准确率分别为 0.98、0.73、0.43、0.84、0.77，而卷积神经网络中未引入 STN 时在上述表情的识别准确率分别为 0.96、0.73、0.45、0.82、0.71。由此表明，在注入旋转样本的前提下，引入 STN 网络在大部分表情中都能比未引入 STN 网络获得更高的准确率，说明在卷积神经网络中引入 STN 网络，能够使卷积神经网络学习到图像的空间信息，使得卷积神经网络能够获得更高的识别准确率。同时，由于中性表情是处于实验室状态下的图像，而在中性表情中引入 STN 网络后并未出现准确率降低的情况，所以说明引入 STN 网络之后不会对不具有旋转角度的图像产生不良影响。

另外，由于本文在卷积神经网络的第一层引入了 STN 网络，相较于未引入 STN 网络的卷积神经网络增加了 32×32×6=6144 个连接，所以本文还针对引入 STN 网络之后所需要的训练时间进行实验。经实验得，引入 STN 网络所需训练时间为 1 383 s，测试时间为 0.229 s，而未引入 STN 所需训练时间为 928 s，测试时间为 0.148 s，由此可见，引入 STN 网络比未引入 STN 网络训练时间多了 455 s，测试时间多了 0.081 s。由此可见，引入 STN 网络虽然增加了一定的训练时间，但是测试时间未明显增加，因此，引入 STN 网络能够在几乎未增加判决时间的基础上有效解决表情识别任务中的旋转不变性问题，从而满足实时处理的需要。

4　结束语

在人脸表情识别任务中，为了充分使用卷积神经网络在训练阶段学习到的分布式特征，并且降低由于局部 ROI 图像包含信息量较少导致的误判对最终判别结果的影响，本文提出了一种二级投票机制对表情图像进行辅助判别方法，并将本文提出的方法与 ROI-KNN 方法以及仅在训练采用 ROI 图像进行数据增强的方法进行比较。实验结果表明，本文提出的二级投票机制能够获得更好的效果。另外，本文还对如何让卷积神经网络学习到表情图像的空间位置信息进行研究，将 STN 网络引入到表情识别任务中，使卷积神经网络具有旋转不变性，提升了系统的鲁棒性。本文提出的方法虽然提升了表情识别的准确率，但是对不同光照、不同角度等复杂情况下的表情图像处理能力还不够强，因此建立一个能在复杂情况下准确识别表情的表情识别系统将是下一步的研究重点。

参考文献：

[1] Hernandez-Matamoros A, Bonarini A, Escamilla-Hernandez E, *et al.* Facial expression recognition with automatic segmentation of face regions using a fuzzy based classification approach [J]. Knowledge-Base Systems, 2016, 110: 1-14.

[2] Shan Caifeng, Gong Shaogang, McOwan P W. Facial expression recognition based on Local Binary Patterns: A comprehensive study [J]. Image and Vision Computing, 2009, 6 (27): 803-816.

[3] 罗源，张灵，陈云华，等. 基于层次结构化字典学习的人脸表情识别[J]. 计算机应用研究, 2017, 34 (11): 3514-3517. (Luo Yuan, Zhang Ling, Chen Yunhua, *et al.* Facial expression recognition based on hierarchy structured dictionary learning [J]. Application Research of Computer, 2017, 34 (11): 3514-3517.)

[4] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C]// Proc of International Conference on Neural Information Processing Systems. 2012: 1097-1105.

[5] Shin M, Kim M, Kwon D S. Baseline CNN structure analysis for facial expression recognition [C]// Proc of the 25th IEEE International Symposium on Robot and Human Interactive Communication. 2016: 724-729.

[6] Perveen N, Singh D, Mohan C K. Spontaneous facial expression recognition: a part based approach [C]// Proc of IEEE International Conference on Machine Learning and Applications. 2017: 819-824.

[7] Hamester D, Barros P, Wermter S. Face expression recognition with a 2-channel convolutional neural network [C]// Proc of International Joint Conference on Neural Networks. 2015: 1-8.

[8] Liu Yize, Chen Yixiang. Recognition of facial expression based on CNN-CBP features [C]// Proc of the 29th Chinese Control And Decision Conference. 2017: 2139-2145.

[9] Meng Zibo, Liu Ping, Cai Jie, *et al.* Identity-aware convolutional neural network for facial expression recognition [C]// Proc of the 12th International Conference on Automatic Face & Gesture Recognition. 2017: 558-565.

[10] Vo D M, Sugimoto A, Le T H. Facial expression recognition by re-ranking with global and local generic features [C]// Proc of the 23rd International Conference on Pattern Recognition. 2017: 4118-4123.

[11] 孙晓，潘汀，任福继. 基于ROI-KNN卷积神经网络的面部表情识别 [J]. 自动化学报, 2016, 42 (6): 883-890. (Sun Xiao, Pan Ting, Ren Fuji. Facial expression recognition using ROI-KNN deep convolutional neural networks [J]. Acta Automatica Sinica, 2016, 42 (6): 883-890.)

[12] Jaderberg M, Simonyan K, Zisserman A, *et al.* Spatial transformer networks [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2015: 665-673.

[13] Lucey P, Cohn J F, Kanade T, *et al.* The extended cohn-kanade dataset (CK+): a complete dataset for action unit and emotion-specied expression [C]// Proc of Computer Vision and Pattern Recognition Workshops. 2010: 94-101.

[14] Hinton G E, Srivastava N, Krizhevsky A, et all. Improving neural networks by preventing coadaptation of feature detectors [J]. Computer Science, 2012, 3 (4): 212-223.

[15] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks [C]// Proc of the International Conference on Artificial Intelligence and Statistics. 2012: 315-323.